DIGIRES

Baltic Research Foundation
for Digital Resilience

Contract number: LC-01682259

D2.1. Report on disinformation detection methodology development

DIGIRES

D2.1. Report on disinformation detection methodology development

Action No:	Connect/2020/5464403
Project and Grant No:	Small-scale online media pilot project DIGIRES; grant number: LC-01682259
Project Title and Acronym:	Supporting Collaborative Partnerships for Digital Resilience and Capacity Building in the Times of Disinfodemic/COVID-19 (DIGIRES)
Deliverable Title:	D2.1. Report on NLP/DL/ML based disinformation detection methodology development
Brief Description:	The document gives the detailed report of activities and results achieved in the framework of the task D2.1. of the project DIGIRES.
WP Number and Task No:	WP2. Develop methodologies for disinformation detection and disclosure and share good practices with EDMO (M03-M15); Task 2.1. Disinformation detection methodology development (M03-M15)
Authors of the Deliverable:	Darius Amilevičius, Andrius Utka, Aistė Meidutė
Contributors:	Vytautas Magnus University, DELFI Lietuva
Nature of the Deliverable:	**R – Report;** P – Prototype; D – Demonstrator; O – Other
Dissemination Level and Audience:	**PU – Public access**; RE – Restricted to other programs or a group specified by the consortium; CO – Confidential

Version	**Date**	**Modified by**	**Comments**
Draft version	January	Darius Amilevičius	The report structure was designed and

European Commission

This project has received funding from the European Union under Contract number LC-01682259

	2023		the key concepts and conceptual ideas illustrating and guiding their practical application were introduced.
1st version	January 17, 2023	Darius Amilevičius Andrius Utka	Necessary clarifications were added.
2nd version	February 13, 2023	Darius Amilevičius Andrius Utka Aistė Meidutė	Conceptual links with other reports, such as the Deliverable **D2.2. Report on disinformation disclosure methodology development** were identified and relevant information was added.
Final version	February 23, 2023	Andrius Utka, Auksė Balčytienė	Final revisions performed, language editing and style corrections added

European Commission

This project has received funding from the European Union under Contract number LC-01682259

Executive Summary

In the framework of the task D2.1 of the project DIGIRES, during the first 3 to 5 months, the group had performed activities aimed at analysis of existing algorithms, methodology, available datasets, and scientific literature. The collected information had allowed the group to identify problems and challenges for disinformation detection methodology development for the Lithuanian language. In parallel, the group had coordinated the cooperation with activities of the project's fact-checker group (see Deliverable **D2.2. Report on disinformation disclosure methodology development**), so that relevant Lithuanian disinformation sources could be identified for experimental testing activities.

The next 6 months of the project had been dedicated to the crawling, accumulation, selection and annotation of the relevant textual desinformation material. As the SOTA advancement of disinformation detection is achieved for the morphologically-poor English language, the group had to produce the comparative analysis between morphologically-rich (Lithuanian) and morphologically-poor (English) languages, so that to assess the usefulness of methods and algorithms. During this period, the group has also experimented with feature extraction solutions (the detailed results of the experiments are presented below).

The last 3 months of the project had been spent developing ML and DL solutions for automatic disinformation detection. Datasets and models compiled during earlier project stages had been used for these tasks. At the end of the project, the group had developed a complex methodological Framework for automatic disinformation detection.

The group of the task had been actively involved in dissemination activities: during the project, the group had kept informing the project's team and scientific community about the progress and results of this task. The group had also prepared the material for a scientific publication. The compiled dataset COVID-19 CORPUS was deposited to the repository of research infrastructure CLARIN-LT for public access (http://hdl.handle.net/20.500.11821/53). The exemplary Framework will be shared via the project's website (https://digires.lt/).

European Commission

This project has received funding from the European Union under Contract number LC-01682259

Table of Contents

European Commission

This project has received funding from the European Union under Contract number LC-01682259

1. Introduction

Disinformation on various media sources is becoming increasingly widespread and it is causing serious concern within the society due to its ability to cause political and social damage with destructive impact. Disinformation and fake news is not only a matter of politicians and regulatory bodies, but it has also become an object of focus for researchers and scientists. In the DIGIRES project, too, we have seen that science can (and must) provide good advice to policy makers. The social sciences and humanities analyse contextual features and can provide new recommendations on how to create an open digital space by empowering groups of people. Meanwhile, technological sciences and innovative machine analysis techniques can be applied to answer questions such as how to create a secure information environment.

The issue of disinformation is also connected to the concept of open science. Open science is one possible means to combat the problem of fake news. Science and journalism have the same goal, which is to separate facts from fiction. Open access aims to ensure that authentic research is distributed as widelly as possible, without much cost since scientific knowledge must be made accessible to all. Open access is not just about disseminating large chunks of data, but also sharing tools and techniques. Research data should benefit the public as well as the scientific community.

By sharing research results on its public events and opening datasets the DIGIRES project is also adhering to main principles of open science: it provides a good example of how science can benefit researchers from various disciplines. The project highlights the need for data sharing and stresses the importance of collaboration between linguists, computer scientists and journalists. All this leads to better research results and furthers the mission of open science.

An extensive research on disinformation detection has been done for the English language and this is no surprise (Jones et al., 2022). English, being the world's lingua franca, has become a major tool for spreading disinformation across different media sources, and, therefore, researchers have access to plenty of useful examples and data for experimentation and creation of disinformation detection methods and tools.

It is quite a different matter with low-resourced languages such as the Lithuanian language: there is a constant lack of resources, including human, data, and tools. The current task on disinformation detection methodology development within the DIGIRES project had been dedicated to improve the situation for the Lithuanian language analysis. This had been done on three levels: 1) theoretical grounding; 2) data compilation; and 3) creation of disinformation detection models capable of identifying suspicious texts.

European Commission

This project has received funding from the European Union under Contract number LC-01682259

Besides, the Lithuanian language is a synthetic language, which considerably differs from analytical languages such as English. This characteristic of the language creates additional challenges along with the lack of resources. Synthetic languages are morphologically-rich and therefore have a greater variety of wordforms than analytical languages. As a result, word-centred models that work well for analytical languages are not always suitable for synthetic languages. The present research tries to overcome the problem by selecting the relevant frameworks and tools.

The following sections of this Report outline outcomes of this task.

European Commission

This project has received funding from the European Union under Contract number LC-01682259

2. Disinformation detection methodology development

Fake news detection is a complex task which requires a multi-faceted approach to address the challenges involved the production of maliscious content (Oshikawa et al. 2020, Rafique et al. 2022). Related problems with the fake news detection task:

1. Fact-checking is the task of assessing the truthfulness of claims made by public figures such as politicians, pundits, etc. Many researchers do not distinguish fake news detection and fact-checking since both of them are to assess the truthfulness of claims. Generally, fake news detection usually focuses on news events while fact-checking is broader.
2. Rumour detection. Rumour detection is the process of identifying and verifying the credibility of informal information that is often spread as user-generated comments via websites, social media, public forums, and blogs. It involves analyzing texts, images, and other media to determine the origin of the information and whether it is credible or not (see also Hangloo et al. 2021).
3. Rumours must contain information that can be verified rather than subjective opinions or feelings.
4. Stance detection is the task of assessing what side of debate an author is on from text. It is different from fake news detection in that it is not for veracity but consistency. Stance detection can be only a subtask of fake news detection since it can be applied to searching documents for evidence.

Our general goal is fake news detection, that is to identify fake news, defined as the false stories that appear to be news, including rumours judged as information that can be verified in rumour detection.

In our research, we focus on fake news detection of text content. As input we take: a) entire articles and b) their titles. There are different types of labeling. We have chosen the binary labeling strategy: “0” (=real) or “1” (=fake). In most studies, fake news detection is formulated as a classification or regression problem. We agree that categorizing all the news into two classes (fake or real) is difficult because there are cases where the news is partially real and partially fake. Common practice is to add additional classes. In our case, besides binary classification we use probability measures. We treat fake news detection as a classification problem (i. e. classifying texts into *real* and *fake*). One of the conditions for fake news classifiers to achieve good performances is to have sufficient labeled data. However, to obtain reliable labels requires a lot of time and labor.

European Commission

This project has received funding from the European Union under Contract number LC-01682259

For research purposes five datasets were created:

1. *DIGIRES COVID-19 Corpus v.1*[1] consists of 351 media articles about COVID-19 pandemics. The corpus was compiled from various internet public Lithuanian media sources. Each article consists of a title and an article body. Corpus contains 351 files in plain text format (TXT) with UTF-8 encoding. In this research the corpus was used for computational linguistics research purposes, e. g. for TF-IDF indexing and as the basis for *DIGIRES COVID-19 ML Dataset v.1* (described below).
2. *DIGIRES COVID-19 ML Dataset v.1*[2]. Manually annotated *DIGIRES COVID-19 Corpus v.1* (described above). 50% of the corpus contains fake news, and 50% - real news articles. Corpus includes three fields: “title”, “text”, and “label”. Sources of articles were selected by professional fact checkers. The annotation of articles was made by two professional fact checkers (see DIGIRES D2.2 Report). Each article contains labels “0” (=real) or “1” (=fake). This corpus was used for TF-IDF[3] vectorisation (described below) and for training of neural networks. Since the corpus is small, the training set and testing set has a ratio 90:10. It is a unique data set, the first of its kind for the Lithuanian language.
3. *Lithuanian media news corpus (ARTICLES-DIGIRES_v1)*. Consisting of c. 500,000 words. Articles collected from various internet public Lithuanian media sources. Mostly from www.delfi.lt. Due to the limitation of publishing rights cannot be shared as open access data. In this research it was used for computational linguistics research purposes, e. g. for generating a word embedding model, creation of neural language model (both described below).
4. *Lithuanian media Fasttext*[4] *word embedding model (ARTICLES-DIGIRES-FAST_v1)*. *Lithuanian media news corpus* was used for creation of this model. The Fasttext algorithm was chosen over traditional word2vec, because word2vec[5] tokenizer as atomic unit uses word level, that is not suitable for the morphologically-rich Lithuanian

[1] Amilevičius, Darius; Utka, Andrius; Meidutė, Aistė and Ruzaitė, Jūratė, 2023, DIGIRES COVID-19 Corpus v.1, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/53.

[2] Amilevičius, Darius; Utka, Andrius; Meidutė, Aistė and Ruzaitė, Jūratė, 2023, DIGIRES COVID-19 ML Dataset v.1, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/54.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

[4] https://fasttext.cc/

[5] https://code.google.com/archive/p/word2vec/

European Commission
This project has received funding from the European Union under Contract number LC-01682259

language[6]. The Fasttext tokenizer uses subwords, so it permits dealing with all morphological forms and in a morphologically-rich language avoids "out-of-vocabulary" problems. This model was used to embed the text and pass to input neuron layer for training and experiments with deep learning algorithms (described below). It is a unique data set, the first of its kind for the Lithuanian language.

5. *Lithuanian media RoBERTa[7] language model (ARTICLES-DIGIRES-ROB_v1)*. The model was created using transformers and the *Lithuanian media news corpus*, consisting of c. 500,000 words. RoBERTa was chosen over BERT[8], because the RoBERTa tokenizer at the atomic level takes symbols, instead the BERT tokenizer at the atomic level takes words. RoBERTa for morphologically-rich languages permits dealing with all inflectional forms and overcomes "out-of-vocabulary" problems. This model was used as an encoder in experiments with Transformer technologies, a fine tuning decoder part for fake news detection tasks (described below). It is a unique data set, the first of its kind for the Lithuanian language.

Our methods for fake news detection experiments are: machine learning models, neural networks models, and computational linguistics.

For baseline evaluation, we have used the corpus DIGIRES COVID-19 Corpus v.1 and machine learning models, and have performed classification using several supervised learning methods, including Multinomial Naive Bayes Classifier, Support Vector Machine, Random Forest Classifier, Gaussian Naive Bayes Classifier, XGBoosts classifier, the input we have vectorized with TF-IDF method (more about the methods in Suhasini and Vimala 2021, Najar et al. 2019, Lilleberg 2015, Do 2021, Ahmed et al. 2020, Piskorski and Jacquet 2020). To achieve best results, min 3-gram and max 5-gram were used to obtain good knowledge of domain lexicon and lexicon usage patterns. First to create the TF-IDF index, the corpus was lemmatised. In our experiment with such a small training data set, all ML algorithms performed very well, achieving more than 90 percent accuracy. Best of all has performed Support Vector Machine Classifier (94 percent accuracy). TF-IDF has provided global domain knowledge (corpus based). Other vectorisation techniques provide only local knowledge (article based). But this method has its shortcomings: out-of-vocabulary problem.

[6] Morphological analyzer (tagger) for Lithuanian has a vocabulary of 185,000 principal word forms (lemmas). We have synthesized c. 20,000,000 morphological forms from this amount based on Lithuanian grammar rules.

[7] https://huggingface.co/docs/transformers/model_doc/roberta

[8] https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

European Commission

This project has received funding from the European Union under Contract number LC-01682259

We can conclude that the baseline method is very suitable with a small training data set and is very useful to predict some insights from a domain general knowledge point of view: a specific text needs to be pointed out to a human review. But it does not suffice to make the final decision.

For medium line evaluation, we have used deep learning method: Long-Short-Term-Memory neural network (to avoid memory of context lost[9]) was used (see also Zhou 2022, Najar et al. 2019), embedding to the input layer was executed: 1) TF-IDF vectorization (Ali et al. 2022) and 2) with the data set Nr. 3. Long-Short-Term-Memory permits to solve gradient vanishing problems (maintain long context) and provide long text to an input layer. With such a small training data set TF-IDF vectorisation permits obtain 84 percent accuracy, because prediction provides global level knowledge variables (e.g. Wang et al. 2022). In the second case, the accuracy dropped down to 56 percent, because Fasttext vectorisation permits to avoid out-of-vocabulary problems, but provides only local level variables.

For advanced line evaluation (SOTA), we have used deep learning method based on Transformers[10] technology and discriminative model (data set Nr. 3) for the encoder part. For the decoder part - fine tuning for fake news detection task - data set Nr. 2 was used. Huge neural language model provides good general language knowledge, but such a small training dataset (to fine tune decoder for specific tasks) does not suffice. For this reason, we obtain only 52 percent accuracy. We have noticed that a transformer-based neural language model provides good general knowledge of language at local level and does not suffer from the out-of-vocabulary problem, but has no global domain knowledge, especially on factual information. To obtain factual knowledge, an external knowledge base must be provided. Another weakness of this solution is the limited input amount. The transformers-based input is limited only to 512 tokens (not words). Many articles are much longer. For this reason, they must be split into sub-articles and in this way the general context of an article may be lost. So this method is most useful for short articles, when there is a large dataset available for training (at least a few thousands of samples).

Based on scientific literature analysis and experiments of our research, we concluded that only machine/deep learning methods for fake news detection are not enough (). As mentioned above, to train a classifier we must obtain reliable labels that require a lot of time and labor. For this reason we can only obtain fake news classifiers with delay in time (after fake news is spread on the Internet). Hence, an ensemble approach must be developed to permit effective

[9] News articles are quite large and due to the inflectional nature of the Lithuanian language and free word order, main parts of sentences have no fixed place and can be separated by long strings of text.

[10] https://huggingface.co/docs/transformers/index.

European Commission

This project has received funding from the European Union under Contract number LC-01682259

feature extraction to identify fake news (Hakak et al. 2020). This ensemble approach consists of:

1. Machine/deep learning classifier for fake news detection in the content of articles (described above).
2. Analyser of article title. Due to the huge amount of daily information, the vast majority of readers read only titles of articles (Horne and Adali 2017). For this reason, fake news creators use various techniques to attract the attention of the reader (Aldwairi and Alwahedi 2018). One of them: emotionally charged long titles.
3. To achieve an ensemble approach goal and create a single powerful predictive model, we must also extract Content-based features: Linguistic, Style, and Semantic. Linguistic-based features aim at capturing the overall intricacy of the news, both in the sentence and word level. They are morphology level features. Part of speech feature distributions can be calculated by using POS tools. POS tools analyse the basic grammar of a text and 'label' it with the appropriate parts of speech. Some studies show that particular distributions of parts of speech may reflect different text functions and might be important for text classification tasks (Rittman et al. 2005, Okulicz-Kozaryn 2013). Style-based features use NLP techniques to extract grammatical information from each document, understanding its syntax and text style. Sentence level features to quantify a reading difficulty score (Carrasco-Farre 2022). Complexity metrics are inspired by readability indexes, such as Simple Measure of Gobbledygook (SMOG), Grade and Automated Readability Index (Reyes and Palafox 2019, Santos et al. 2020), which use words and by counting the number of unique words divided by the total number of words, measuring the vocabulary variation of the document. Those textual statistics are intended to help the characterization of the complexity of differences between news classes. Type-Token Ratio (TTR) is also extracted. The Stylometric features take into account more advanced NLP techniques to extract grammatical and semantic characteristics from the text. Both of them use computational linguistic methods and are content based. Semantic-based features use NER and sentiment analysis techniques to extract entities (persons, locations etc.) and classify emotion level in the text (Alonso et al. 2021).
4. Network-based features (also known as Social context based) Social proofs are based on the network analysis and are external to content. They show public reaction and the level of participation in a discussion or personal/group attitudes towards a message. Social proofs also contain some attitudes of the audience towards the source of a message (by some they are considered as bad, by others as good). In our view, it is inappropriate to add the attribute of pre-assessment of the source ("fake-source" or "good-source"), as this creates the potential for discrimination: any source

This project has received funding from the European Union under Contract number LC-01682259

may contain truthful information alongside misleading information (e.g. Raza et al. 2022).

To realize the feature extraction system, we have developed a morphological analyzer included in SpaCy[11] framework and have developed tools for content-based and stylistic-based feature extraction. For NER recognition we have used a standard NER solution, included in the SpaCy framework. Sentiment analyzer and network-based feature extractors were not developed in this project.

All solutions, described above, are connected together in a Framework, that is a single predictive model to recognize fake news.

Due to the small amount of dataset for training, it must be considered as a prototype. This prototype produces a recommendation that some articles must be reviewed by a fact checker because there is something wrong in it. A prototype produces general evaluation of the article in consideration (probability) and detailed evaluation of all parameters (classifier prediction, by all features etc.). Our solution is scalable: it can be expanded onto other morphologically rich languages (more about methogology in Abonizio et al. 2020) and other topics of fake news. Main weakness of our prototype is that it is more reactive than proactive.

[11] spacy.io.

European Commission

This project has received funding from the European Union under Contract number LC-01682259

3. Further development of our prototype

Our research has demonstrated that several machine learning techniques are effective in detecting false information; it also has shown that ensamble (or hybrid) models combining several techniques have the potential to further improve the performance of disinformation detection. Thus, further steps are required to improve the accuracy of these models:

1. Further development of training data set (number of samples and more topics).
2. Creation of sentiment analyzer suitable for emotions in news evaluation.
3. Development of syntactic analyser (parser) for Lithuanian that permits the development of rhetorical analysis of articles.
4. Development of tools for network-based features extraction.
5. Research for additional tools to make our solution more proactive.
6. Development of automatic extraction of articles from various internet sources
7. Add the prediction model for identifying texts generated by GPT-based tools.
8. Using of GPT3-based solutions for detecting whether a certain piece of text is generated by a human or by GPT3-based text generator. (GPT3-based generators produce texts that are very similar to texts generated by humans. Presently, such generated texts can only be detected with the use of the same model (e.g. GPT3) that has generated it.)

European Commission

This project has received funding from the European Union under Contract number LC-01682259

4. Conclusions

In the context of our research and scientific conclusions, we can draw the following more general conclusions about using SOTA technological tools and methods for the automatic fake news combating/detecting in the DIGIRES project.

Prior to giving the answer to the question “Can we use AI to identify fake news?”, we must answer a very nuanced question: how do we know “what is true, and what is false?”. The category of truth belongs to the realm of knowledge and ethical domain. Truth is the value that humans hold and use to challenge other views. Truth can also include belief. Truth is a moving target for an individual. Still, there’s a fluid set of local and global truths all the time. There are a lot of truths in between those that are more difficult to classify. Truth varies in space and time. Rather than tackling this question head-on, researchers try to answer simpler variants of it and for this purpose make use of technologies:

1. Content-based approach. SOTA AI operates at the “keyword” level, flagging words and word patterns and looking for statistical correlations among them and their sources. This can be somewhat useful: statistically speaking, certain patterns of language may indeed be associated with dubious stories. But none of such correlations reliably sort the true from the false. What makes the article “mostly false” is that it implies a causal connection that is not always directly expressed in word patterns. Causal relationships are where contemporary machine learning techniques start to stumble. Understanding the significance of the article in consideration also requires understanding of multiple viewpoints. Most current AI systems that process language are oriented around a different set of problems. AI systems that have been built to comprehend news accounts are extremely limited in multi-problem solving and rarely go much further, lacking a robust mechanism for drawing inferences or a way of connecting to a body of broader knowledge. The results of our research (also of others) argue that AI cannot fundamentally tell what’s true or false — this is a skill much better suited to humans. To become more autonomous in fake news identification, AI will require the development of a fundamentally new AI paradigm, one in which the goal is not to detect statistical trends, but to uncover ideas and the relations between them. Doing so would require a number of major advances in AI taking us far beyond what has so far been invented. For now, a statistical model “what is true” is injected to AI systems in the form of a human-annotated training dataset. That is, its starting point is subjective-human based (determined by opinion of fact checkers), but not based on AI cognitive capabilities (e.g. Yu et al. 2003). For this reason, the final decision about the veracity of an article must be the responsibility of a human analyst.

European Commission

This project has received funding from the European Union under Contract number LC-01682259

2. Spreading prevention approach. Automatic fake news generation bots contribute to mass and fast spreading of fake news. In the past, techniques of analyzing linguistic cues such as word patterns, syntax constructions, readability features, etc. were used to differentiate (up to a certain accuracy) between human and computer-produced text. But OpenAI's recently launched GPT-3 can write essays, stories, emails, poems, business memos, technical manuals etc. It can answer philosophical questions, simplify legal documents and translate between languages. It is almost impossible to separate GPT-3's output from a human-written text in short pieces of text. In this case, only GPT-3 based AI tools can determine whether a piece of content was created by a computer or a human. For this reason AI can be used as an aiding tool in the generation of fake news at a big scale. To determine whether a piece of content was created by a computer or a human and verify its veracity more complex tools set must be used.
3. Private messaging platforms. Due to the nature of these platforms that are producing enormous constant streaming traffic of information (Facebook: 1.7m pieces of content per minute; Twitter: 347.2K tweets per minute[12]) with access restrictions (e.g. WhatsApp supports end-to-end encryption), it becomes almost impossible to monitor the communication on private messaging platforms to fight fake news.
4. The social sciences and humanities analyse contextual features and can provide new recommendations on how to create an inclusive and open digital space by empowering different groups of people, whereas technological sciences and innovative machine analysis techniques can be applied to answer questions such as how to create a secure information environment. In such a way (and by combining both approaches), a digitally sustainable information space might be created which is both efficient and effective to reach its (discoursive and deliberative) outcomes[13] (Jaramillo et al. 2020). From the technological point, such a space will be more secure and accessible, and from the deliberative – giving users greater control of the information they are accessing.
5. The issue of disinformation is also connected to the concept of open science. Open science is one possible means to combat the problem of fake news. Science and journalism have the same goal, which is to separate facts from fiction. Open access aims to ensure that authentic research is distributed as widely and openly as possible.

[12] https://www.domo.com/data-never-sleeps#top

[13] See Deliverable **D1.5. Media policy suggestions** and the Policy Brief 'Informed Deliberation and the Digital Age: A Question of Quality of Media Texts', which provides conceptualization for the Deliberation Quality Index (DQI).

European Commission

This project has received funding from the European Union under Contract number LC-01682259

Automated indicators (presented in this Report) are not final, they are just the start of collective learning. Researchers could provide timely and relevant indicators, but the readers of these results probably know the local context better, and can better interpret the results. This co-learning combined with a trial-and-error process will strengthen the resilience of information governance as a collective capacity.

6. As further hypothesized, Online Deliberation research[14] should combine multiple methods to investigate different aspects of the same empirical phenomenon. Future research should focus on developing automated indicators by combining natural language processing, network analysis, time-series analysis, and other methods. In the light of recently emerging SOTA large language models (GPT4, Bard, etc.) developed by large private companies (Google, OpenAI/Microsoft, etc.), current automated computational methods for assessing online deliberative quality must be revised in the near future (see also Zelers 2020). The revision should be done in close collaboration between researchers and policy makers. SOTA LLMs are capable of mimicking humans and producing high quality coherent texts at large scale in seconds, besides LLMs are not accessible to researchers for close analysis. Recommendations for policy makers: 1) release revised AI Act that regulates private companies that are in possession of LLMs; 2) oblige private companies to make access to analysis of LLMs to researchers; 3) provide appropriate funding for such kind of research, as it requires huge amount of computational power, data, and human resources.

To conclude, the small scale pilot project (SMPP) DIGIRES was beneficial to us in several aspects. Conceptualizations of digital (false) information must be viewed not only from a position that innovative technologies are a valuable driver of scientific experimentation. Rather, science must focus on providing policy makers with evidence-based guidelines for strategic recommendations. By testing out interdisciplinary approaches and application of analysis techniques, the DIGIRES project demonstrates how a combination of conceptual approaches and innovative machine learning analysis methods can be utilized to inform about specific (quantitative and qualitative) features of the digital content (see Deliverable **D1.5. Media policy suggestions – DQI: deliberation quality identification).**

[14] See the Policy Brief 'Informed Deliberation and the Digital Age: A Question of Quality of Media Texts', which provides conceptualization for the Deliberation Quality Index (DQI), which is provided together with Deliverable **D1.5. Media policy suggestions.**

European Commission

This project has received funding from the European Union under Contract number LC-01682259

In 2023, the experimentation with the DQI idea (deliberation quality identification assessment) will be further extended into the next stage, namelly the Tasks of the **BECID Hub project**[15], which has started in December 2022.

[15] https://becid.ut.ee

European Commission

This project has received funding from the European Union under Contract number LC-01682259

References

Ahmed, Alim Al Ayub, Ayman Aljarbouh, Praveen Kumar Donepudi, Myung Suh Choi, Detecting Fake News using Machine Learning: A Systematic Literature Review. https://arxiv.org/abs/2102.04458. 2021.Ray Oshikawa et alias, A Survey on Natural Language Processing for Fake News Detection. 2020.

Aldwairi, Monther and Ali Alwahedi. Detecting Fake News in Social Media Networks. The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018). 2018.

Alonso, Miguel A. et alias, Sentiment Analysis for Fake News Detection. 2021. *Electronics*, *10*(11), 1348; https://doi.org/10.3390/electronics10111348. 2021.

Carrasco-Farré, Carlos, The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. https://doi.org/10.1057/s41599-022-01174-9 . HUMANITIES AND SOCIAL SCIENCES COMMUNICATIONS. 2022.

Do, Tien Huu et alias, Context-Aware Deep Markov Random Fields for Fake News Detection. Published in: IEEE Access (Volume: 9). DOI: 10.1109/ACCESS.2021.3113877 . 2021.

Hakak, Saqib et alias, An ensemble machine learning approach through effective feature extraction to classify fake news. doi:10.1016/j.future.2020.11.022. 2021.

Hangloo, Sakshini et alias, Fake News Detection Tools And Methods – A Review. arXiv:2012.11185[cs.CY]. 2021.

Horne, Benjamin D. and Sibel Adalı, This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. DOI: 10.1609/icwsm.v11i1.14976 . 2017.

Jaramillo, Maria Clara and Jürg Steiner, From discourse quality index to deliberative transformative moments. In *Handbook of Democratic Innovation and Governance*. Publisher: Edward Elgar. 2020.

Jones, Keenan, Enes Altuncu, Virginia N. L. Franqueira, YHichao Wang and Shujun Li, A Comprehensive Survey of Natural Language Generation Advances from the Perspective of Digital Deception. https://arxiv.org/pdf/2208.05757.pdf , 2022.

Lilleberg, Joseph et alias, Support Vector Machines and Word2vec for Text Classification with Semantic Features. Proc. 20151IEE 14th Inl'l Coni. on Cognitive Informatics & Cognitive Computing IIccrcn51. N. Ge. I.lu. Y. Wang. N. Howard. P. Chen. X. Tao. B. Zhang. & LA. Zadeh IEdsJ 918-1-4613-1290-91151$31.00, IEEE. 2015.

This project has received funding from the European Union under Contract number LC-01682259

Marish Ali, Abdullah et alias, Deep Ensemble Fake News Detection Model Using Sequential Deep Learning Technique. 2022. https://www.mdpi.com/1424-8220/22/18/6970

Najar, Fatma et alia, Fake News Detection using Bayesian Inference, 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). 2019.

Nath, Keshav, Priyansh Soni, Anjum, Aman Ahuja, Rahul Katarya, Study of Fake News Detection using Machine Learning and Deep Learning Classification Methods. Conference: 2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT). 2021.

Okulicz-Kozaryn, Adam, Cluttered writing: adjectives and adverbs in academia. Scientometrics 96: 679-681. DOI 10.1007/s11192-012-0937-9. 2013

Oshikawa, Ray et alias, A Survey on Natural Language Processing for Fake News Detection. LREC, 2020.

Piskorski, Jakub and Guillaume Jacquet, TF-IDF Character N-grams versusWord Embedding-based Models for Fine-grained Event Classification: A Preliminary study. Proceedings of AESPEN 2020, pages 26–34. Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020.

Queiroz Abonizio, Hugo et alias, Language-Independent Fake News Detection: English, Portuguese, and Spanish Mutual Features. Future Internet 12, 87; doi:10.3390/fi12050087. 2020.

Rafique, Adnan et alias, Comparative analysis of machine learning methods to detect fake news in an Urdu language corpus. DOI 10.7717/peerj-cs.1004. 2022.

Raza, Shaina and Chen Ding, Fake news detection based on news content and social contexts: a transformer-based approach. International Journal of Data Science and Analytics 13:335–362 https://doi.org/10.1007/s41060-021-00302-z . 2022.

Reyes, Jesus and Leon Palafox, Detection of Fake News based on readability. https://openreview.net/forum?id=ByxTOnokxr. 2019.

Rittman, Robert, Nina Wacholder, Paul Kantor, Kwong Bor Ng, Tomek Strzałkowski, and Ying Sun, Adjectives as Indicators of Subjectivity in Documents. In Proceedings of the 67th ASIS&T Annual Meeting, vol 41, 349-359. 2005.

Santos, Roney et alias, Measuring the Impact of Readability Features in Fake News Detection. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1404–1413 Marseille, 11–16 May 2020.

European Commission

This project has received funding from the European Union under Contract number LC-01682259

Suhasini, V. and N. Vimala, A Hybrid TF-IDF and N-Grams Based Feature Extraction Approach for Accurate Detection of Fake News on Twitter Data. Turkish Journal of Computer and Mathematics Education Vol.12 No.06, 5710-5723. 2021,

Wang, Depei et alias, Few-Shot Text Classification with Global–Local Feature Information. https://doi.org/10.3390/. 2022.

Yu, Hong and Vasileios Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. https://aclanthology.org/W03-1017.pdf. 2003.

Zellers, Rowan et alias, Defending Against Neural Fake News. https://openreview.net/pdf?id=HygSC4BlUB . 2020.

Zhou, Hai, Research of Text Classification Based on TF-IDF and CNN-LSTM. Journal of Physics: Conference Series 2171 012021. IOP Publishing. doi:10.1088/1742-6596/2171/1/012021. 2022.

European Commission

This project has received funding from the European Union under Contract number LC-01682259